Proceedings of ANAS (Biological and Medical Sciences), vol. 73, No 2, p. 25-32 (2018)

Outlier Detection in Atomic Temperature Factor - B Value Distribution

R.C. Masmaliyeva[1] and G.N. Murshudov[1,2*]

[1]*Laboratory of Computational Structural Biology, Institute of Molecular Biology & Biotechnologies, Azerbaijan National Academy of Sciences, 2A Matbuat Ave., Baku AZ1073, Azerbaijan*
[2]*MRC Laboratory of Molecular Biology, Cambridge CB2 0QH, UK*
**E-mail: garib@mrc-lmb.cam.ac.uk*

Several methods for outlier detection in Atomic Displacement Parameters – B value is applied to one particular macromolecular structure. It is demonstrated that outliers in B values give good indication errors in the atomic models. In this particular example it is demonstrated that a hypothesis that atomic B values distribution is shifted inverse gamma distribution. Removal of outliers from the set of B values improves the estimation of the parameters of the distribution. Local outliers in B values indicate errors in atomic models: validity of the assumption that neighbouring atoms must have similar B values has been verified. It is expected that local and global outlier detection program and modelling of B values as inverse gamma distribution will help to select the reliable atomic models. We suggest that such outlier detection and modelling should be part of model building and refinement of macromolecular structures using crystallographic diffraction data and single particle cryo electron microscopy maps.

Keywords: *Macromolecules, validation, macro-molecular crystallography, outlier detection, inverse gamma distribution*

1. INTRODUCTION

There are three main methods to derive atomic models of biological macromolecules (Berman et al., 2012; Cavagnero, 2003): crystallography, electron microscopy and nuclear magnetic resonance (NMR) (Rupp, 2010; Frank, 2006; Clayden et al., 2001). Both crystallography and electron microscopy use scattering of particles whereas NMR uses various spectroscopic measurements to derive structural information. Atomic models are derived using software packages that use different assumptions about the nature of experiment as well as molecules under study. The resulting atomic structures should be considered as statistical models and they have to be validated using as independent as possible validation tools (Rupp, 2010; Papageorgiou, Mattsson, 2014; Henderson et al., 2012). Validations must be done against experimental data as well as against prior knowledge about macromolecules. There is a number of software tools dealing with the problem of validation of atomic structures (Chen et al., 2010). In this work we would like to address one of the problems standard validation programs usually ignore – validation of atomic displacement parameters (ADP). After all, if atomic displacement parameters can indicate the level of accuracy of atomic positions, validation of atomic positions and interatomic distances should be adapted accordingly; there is no point of validating wrong atoms against chemical and structural information.

In crystallography and electron microscopy studies of molecules observed densities are modelled as a sum of Gaussians centred at the atomic positions (Rupp, 2010):

$$\rho(x) = \sum_{i=1}^{N} \rho_i(x - x_i)$$

with:

$$\rho_i(x) = \sum_{j=1}^{N_{Gauss}} \frac{c_{ii}}{(2\pi(u_i+u_{ij}))^{3/2}} e^{-\frac{|x|^2}{2(u_i+u_{ij})}} \qquad (1)$$

Where u_i is the property of the observed atoms in the molecule, u_{ij} are properties of chemical elements or atom types. Each atomic type is described by N_{gauss} Gaussians, usually N_{gauss} = 5. c_i are weights of Gaussians. x_i are vectors of atomic positions, x is a vector of position in the three-dimensional space. These models, although are approximations to true densities, work sufficiently well in practice. They do not account such fine electronic details as bonding electrons or charge redistribution as a response to interactions with the environment. $\rho(x)$ has different meaning for different scattering methods: if X-rays are used then $\rho(x)$ is the electron density, if electrons are used then it is the electrostatic potential of the molecule. As it can be seen from the formula when u_i becomes large then the density corresponding to this atom become smeared out or blurred. If different atoms have wildly different ADPs then it can be expected that these atoms will have very different densities corresponding to their ADPs. If ADPs would represent only oscillation of atoms around their centre then it would reflect the relative mobility of atoms. In general, it can be

expected that oscillations are different in different directions resulting in anisotropic ADPs. In this work we consider only isotropic ADPs.

Fourier space counterpart of the expression (1) is:

$$F(s) = \sum_{i=1}^{N} f_i(s) e^{-2\pi^2 u_i |s|^2/4} e^{-2\pi \iota s x_i} \quad (2)$$

Where *F(s)* is the Fourier transformation of the density, s is the vector of positions in the Fourier space, f_i is the scattering factor of the atom:

$$f_i(s) = \sum_{j=1}^{N_{gauss}} c_{ij} e^{-2\pi^2 u_{ij} |s|^2}$$

In both representations we essentially assume that the contribution from atoms to the density is convolutions of Gaussians describing atomic mobility or uncertainty and atoms at rest. Uncertainty associated with the Gaussians is called atomic displacement parameters. There are several contributors to the ADP including: 1) dynamic and static disorder in crystals that combine such factors as atomic mobility, crystal lattice disorder: 2) errors in atomic positions, these have the effect of increasing atomic displacement parameters to compensate for errors in the atomic positions; 3) misidentified atom types, this can either increase or decrease ADPs. It would be very difficult to disentangle these contributions without additional information, therefore when dealing and trying to interpret atomic models with the models from the PDB we have to bear this in mind.

Since every atom is linked with the neighbouring atoms either via covalent bond or via non-bonding interactions we can assume that neighbouring atoms have similar oscillation, if they deviate from each other too much then such atoms should be considered as suspicious and they must be revised with care using the observed/estimated density. It should also be mentioned that ADPs of atoms define their relative contribution to the Fourier coefficients via Debye-Waller factor (Debye, 1913) which essentially states that this contribution can be described as a Gaussian in three-dimensional space.

Since ADPs are directly related to the oscillation or uncertainties of the atoms which are modelled as Gaussian that means that ADPs are proportional to the second central moments or variance of the normal probability distribution. In Bayesian statistics it is usual to model the distribution of variances of the normal distribution as an Inverse Gamma distribution (Witkovsky, 2001; Cook, 2008). This distribution has been used successfully as natural conjugate priors for Bayesian modelling of data with Gaussian population distribution (Murphy, 2007).

Since B values are related to the errors in the atomic model they sometimes are used to select reliable set of atoms for further analysis (Chen et al., 2010), therefore it is important to design a procedure that would allow reliable detection of atoms with unusual ADPs; if ADP is too high then it is likely that this atom has been wrongly placed, if it is too small then this atom may have been misidentified, i.e. it might be heavier than that in the PDB file.

In this paper we will describe several outlier detection algorithms for isotropic ADPs for a single entry from the PDB. We will discuss the global and local outliers. We will also demonstrate that removal of outliers improves the estimation of the parameters of shifted inverse gamma distribution proposed as a model for B value distributions.

2. METHODS

There is a number of outlier detection methods described in the literature (Barnett, 1994; Iglewicz, 1993; High, 2000). In this paper we will discuss the methods we managed to use successfully for analyses of ADP distribution.

*Tukey's box and whisker plot method (*Hartwig, 1979): It is a widely used method in descriptive and exploratory statistical data analysis. In this method such statistics as the median, 1st and 3rd quartile, lower and upper extreme values are plotted on the same plot. Such plots help to visually inspect the data and see if there are outliers. Tukey's rule of determining outliers consists of the following steps: 1) calculate the interquartile range as IQR = 3rd quartile– 1st quartile; 2) calculate the upper fence as UpperFence = 3rd quartile + k*IQR; 3) calculate the lower fence as LowerFence = 1st quartile – k*IQR. Points that fall below the lower or above the upper fence are considered outliers. Here k is a factor used to identify outliers with various severities: k=1.5 is used for "mild" and k=3 is used for "extreme" outliers. In our application we need to remove only extreme outliers.

Z-score: Another standard method is Z-score (Shiffler, 1988) that is used to detect outliers in the data with using standard deviation and mean. For each data point Z values are calculated using the formula:

$$Z_i = \frac{x_i - \bar{x}}{sd}$$

where $\bar{x}$ and *sd* are the and standard deviation of the data. Z-scores with an absolute value greater then k are generally considered as outliers. Usually k=3 is taken as default which works in practice sufficiently well. In this paper Z-score method was used with the parameter k=3. This method is not robust to outliers, existence of outliers affects the mean and the standard deviation calculated from the sample. Moreover this method works well with the data points sampled from the population with symmetric distribution.

Modified Z-score: Iglewicz and Hoaglin

(Iglewicz, 1993) recommend using:

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}$$

where $\tilde{x}$ and MAD are the median and the median absolute deviation respectively. MAD is the median of the absolute differences between the data points and the median of the data. These statistics are often used in robust statistical estimations: median replaces the mean and MAD replaces the standard deviation. If the population from which the data have been drawn has the normal distribution then median is equal to the mean and MAD/0.6745 is equal to the standard deviation (Venables, 1999). Although the authors recommend that modified Z-scores with an absolute value of greater than k=3.5 to be considered outliers, in practice to detect outliers with various severity different values of k should be used.

For local analysis first for each atom the list of its neighbours was calculated using the efficient cell algorithm (Mattson, 1999). For this 4.2Å radius was used, although the radius is a tuneable parameter. Then for each atom B values of its neighbourhood was analysed.

3. RESULTS

3.1. Global analyses and outliers. B values of the macromolecular structures are proxies for atomic mobility as well as errors in the model. The modelling of the B value distribution is important for understanding of fundamental properties of positional errors and atomic mobility. They can also be used for outlier detection and in future for map calculation. As it was mentioned in [Masmaliyeva, Murshudov 2017, Dauter 2006], the distribution of B values can be approximated by a shifted Inverse Gamma (IG) distribution. In this paper, as an example we used the protein structure with the PDB code 4XKT (resolution 1.82, R factor 0.17, Rfree 0.19) [Bradley et.al. 2015]. Since it is likely that the structures in the PDB have been under-refined, before any further analysis such as outlier removal and estimation of the parameters of the distribution, the structure was re-refined using the maximum likelihood refinement program Refmac5 (Murshudov et al., 2004) from the CCP4 (Wimm et al., 2011). We applied above described methods to determine outliers and to calculate the parameters of the distribution before and after removal of outliers. The results of the estimations are given on Table 1. Figure 1 illustrates the histogram and the fitted density plot of the initial B value distribution of the protein structure. Figure 1 and Table 1 show that the initial distribution has a long right tail and low shape parameter respectively. As it was mentioned in (Masmaliyeva, Murshudov, 2017) the shape parameter alpha should be around 3.5.

In Tukey's method for detection of “mild” outliers determined with the factor = 1.5 are too sensitive for B value distribution as shown in the Table 1 and Figure 2 (a). As it is described in (High, 2000; Hartwig, Dearing, 1979), for asymmetric distributions, data values below InnerFence are not always outliers and the values higher than OuterFences are almost always outliers.

It is known (Leys et al., 2013) that the methods based on median absolute deviation instead of mean and standard deviation are more robust to outlier methods because median and median absolute deviation themselves are not affected by few outliers in contrast to mean and standard deviation. Median is robust to 50% outliers meaning that its breakdown point is at 50%, MAD is robust to up to 40%. In our application the method using standard Z-score method seems to give more sensible answers. The reason for this will be a part of future detailed analysis.

The number of outliers detected with each method mentioned above is given on Table 1. There are 610 atoms with outlying B value which is detected by all methods mentioned above. In respect that these atoms were detected by all considered methods, we expect them to be true outliers. Figure 5 drawn by the model building program Coot (Emsley, 2010) shows the electron density of two the amino acid residues detected as outliers. It is clear there is no electron density for these atoms indicating that these residues have been modelled incorrectly. There are just 51 B values which determined as an outlier by just one method and all them are results of Tukey’s method with k=1.5 factor. This means that k=1.5 is very low and should be treated carefully.

Table 1. Number of outliers detected with different methods in B value distribution of 4XKT protein

4XKT	Outliers number	B_0	Min	Max	Mean	Median	Variance	Skewness	Kurtosis	1st Q	alpha	beta
Initial distribution	--	5.356	5.94	131.33	15.98	12.35	154.3	4.52	27.65	10.26	2.77	17.89
Tukey’s method (factor=1.5)	1271	4.828	5.94	26.78	13.22	11.96	17.621	1.11	3.64	10.12	4.5	30.23
Tukey’s method (factor=3)	610	4.9	5.94	36.68	13.96	12.16	29.99	1.63	5.64	10.2	3.87	26.45
Z-score method	383	5.01	5.94	53.24	14.39	12.22	42.16	2.22	9.36	10.22	3.54	23.99
Modified Z-score method	819	4.86	5.94	32.66	13.7	12.09	24.76	1.43	4.86	10.17	4.1	27.94

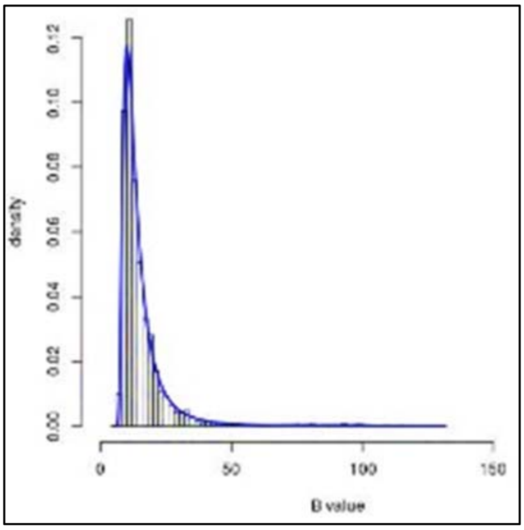


Fig. 1. Initial B value distribution of the protein 4XKT.

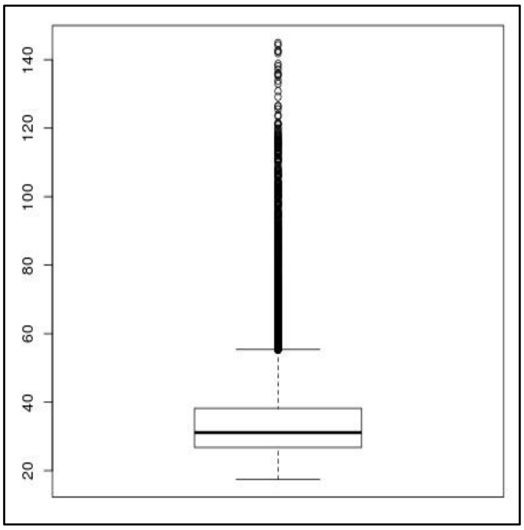
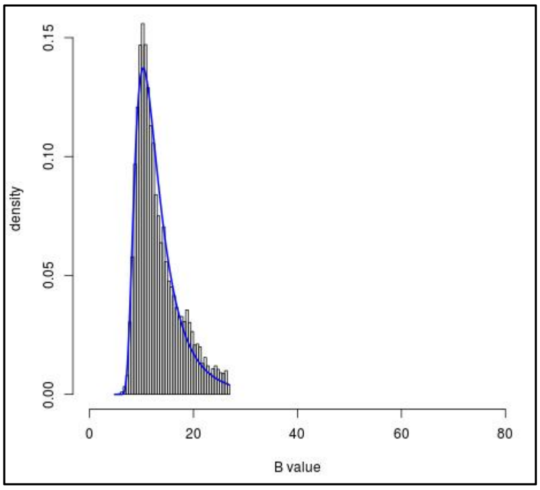
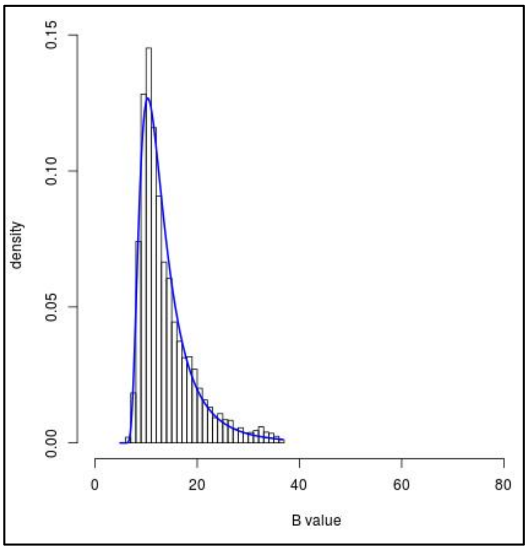


Fig. 3. Box-plot of initial B value distribution of the protein 4XKT.

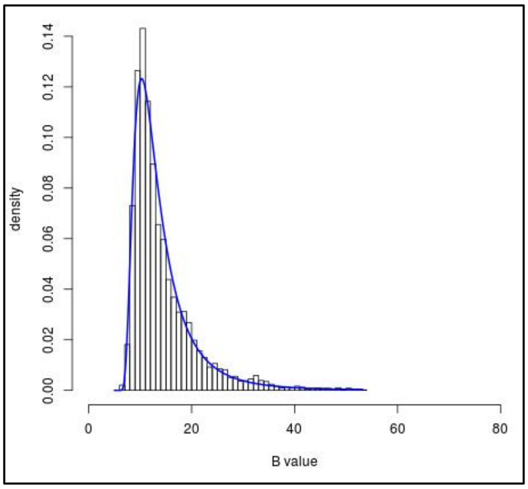


(a)

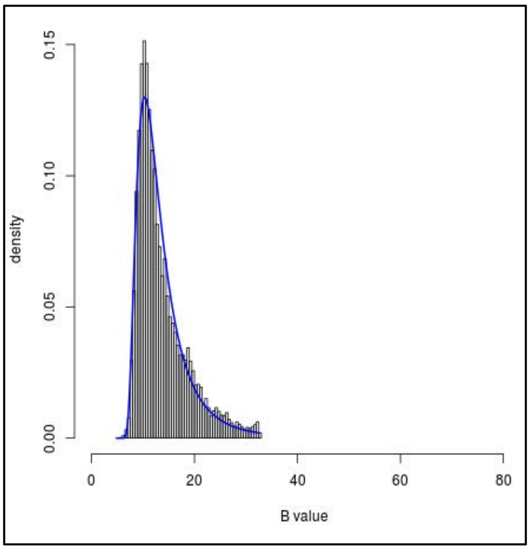


(b)

Fig. 2. B value distribution of the protein 4XKT after removing outliers with Tukey's method with **(a)** k=1.5 and **(b)** k=3.



(a)

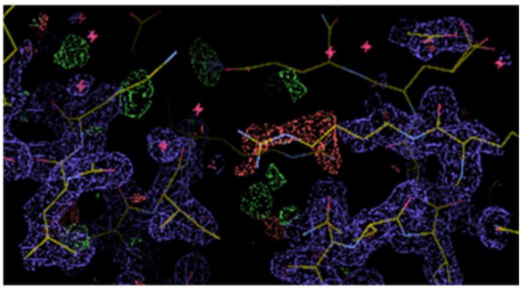


(b)

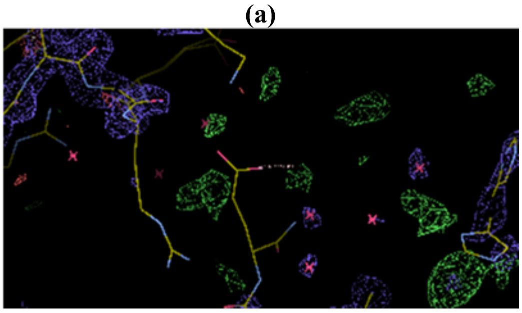
Fig. 4. B value distribution of the protein 4XKT after removing outliers with **(a)** Z-score and **(b)** modified Z-score method

3.2. Local analysis and outlier detection. When atoms are incorrectly placed, the atomic B values become much larger than those of neighbouring atoms, reflecting errors in the model. It is generally expected that neighbouring atoms should have similar B values in regions where modelled atoms are positioned accurately. If neighbouring atom have wildly different B value after refinement, then it usually means that some of the atoms are either 1) in the wrong place; or 2) incorrectly parameterised, for example, occupancies and/or element types for some of the atoms are wrong (Masmaliyeva, Murshudov, 2017).

To detect outliers of atoms in their local environment modified standard deviation was used. As it is mentioned above $sd_{modified} \approx MAD/0.6745$ was used. In the local analysis we detected 4117 atoms with outlying value of B factor value. The largest outlier with B value $98.96 sd_{modified}$ corresponded to OE1 of 157th GLU residue of the chain D (Figure 5 b). In Figure 6 residues with a local outliers described in ball-and-stick mode. With 4.2 radius and modified SD 10.7, 4117 atoms with outlying B value were detected.

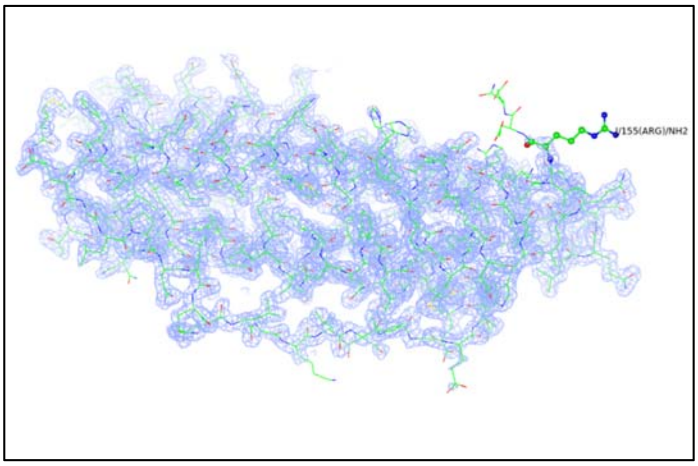


(a)

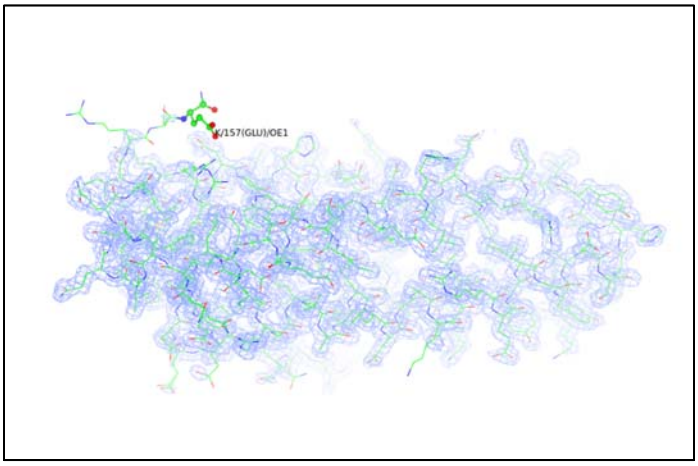


(b)

Fig. 5. Electron density of some residues containing an atom with outlier B value. (a) 39th residue ARG of A chain; (b) 157th GLU residue of D chain. This figure was drawn using coot [Emsley 2010] (Map sigma = 0.343415).



(a)



(b)

Fig. 6. Examples of local outliers; electron density of K and J chains of the protein 4XKT with labelled "outliers".
This figure was drawn using ccp4mg (Nicholas, 2011) (Map sigma = 0.49).

4. CONCLUSION AND FUTURE PERSPECTIVES

Outliers are the data points which strongly deviate from the centre of the distribution. In this paper, global and local B value outliers of three-dimensional structures of macromolecules are discussed. Outliers of B values in a structural model indicate errors and/or misinterpretation of the scattering data during model building and refinement. Several outlier detection techniques have been used. These are Tukey's boxplot, Z-score and modified Z-score methods. Removing the extreme values of B values improves statistical estimation of the B value distribution – shifted inverse gamma distribution. B value outlier detection should be used as a part of model building and refinement. It will ensure that atoms are positioned correctly resulting in more accurate atomic models that are usually used for drug design and bioinformatics analysis purposes.

When B values are larger than that of the rest of the atoms then it means that either these atoms are in wrong place or wrongly parametrized. However, during refinement of atomic models using scattering data it is better to assume that the B-values reflect atomic mobility. Therefore, in such cases it is better to restrain the B-values of neighbour atoms to be similar to each other. If they differ wildly it is usually an indication that model contains errors; these errors should be detected and corrected during modelling stage – if it is done on time and with care then accuracy of the resulting atomic models can be increased substantially.

The results of this paper will in future be implemented in a python language based program and distributed to the structural biology community to help them to correct atomic models during model building and refinement.

In future we also plan to extend of B value analyses for modelling of the distributions and detection of outliers for anisotropic B value cases. It seems that by analogy with the isotropic B value distribution the distribution of anisotropic B values should be modelled using the inverse Wishart distribution [Haff 1979] which is used as conjugate priors for multivariate normal distribution. We will also design new methods for anisotropic B value outlier detection: one potential candidate for this is BACON algorithm (Nedret, 2000) which seems to be able to detect with sufficient accuracy outliers in multivariate data.

ACKNOWLEDGEMENTS

This work was supported by the Presidium of Azerbaijan National Academy of Sciences - grant of decree № 5/9 dated on 15.03.2017. GNM also thanks MRC grant MC_US_A025_0104.

REFERENCES

Barnett V., Lewis, T. (1994) Outliers in statistical data. 3rd ed. Wiley**: 176 p.**

Berman H.M., Battistuz T., Bhat TN., Bluhm WF., Bourne P.E., Burkhardt K., Feng Z., Gilliland G.L., Iype L., Jain S., Fagan P.., Marvin J., Padilla D., Ravichandran V., Schneider B., Thanki N., Weissig H., Westbrook J.D., Zardecki C. (2002) The Protein Data Bank. *Acta Crystallogr. D.,* **58:** 899-907

Bradley J.M., Svistunenko D.A., Lawson T.L., Hemmings A.M., Moore G.R., Le Brun N.E. (2015): Three Aromatic Residues are Required for Electron Transfer during Iron Mineralization in Bacterioferritin. *Angew. Chem. Int. Ed. Engl.,* **54:** 14763-14767.

Cavagnero S. (2003) Using NMR to determine protein structure in solution. *Journal of Chemical Education.* **80 (2):** 125.

Chen V.B., Arendall W.B., Headd J.J., Keedy D.A., Immormino R.M., Kapral G.J., Murray L.W., Richardson J.S., Richardson D.C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica,* **D66:**12-21.

Clayden J., Greeves N., Warren S., Wothers P. (2001) Proton nuclear magnetic resonance. *Organic Chemistry*. Oxford University Press, **Chapter 11:** 269 p.

Cook J.D. (2008) Inverse Gamma Distribution. Tech. Rep. *http://www.johndcook.com/inverse gamma.pdf.*

Dauter Z., Murshudov G.N., Wilson K.S. (2006) Refinement at atomic resolution. *International Tables for Crystallography Volume F: Crystallography of biological macromolecules.* M.G.Rossmann and E.Arnold (Eds.). **Vol. F: Chapter 18.4:** 393-402

Debye P. (1913) Interferenz von Röntgenstrahlen und Wärmebewegung. *Annalen der Physik.*, **348 (1):** 49-92.

Emsley P., Lohkamp B., Scott W., Cowtan K. (2010) Features and Development of Coot. *Acta Cryst.,* **66:** 486-501.

Frank J. (2006) Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state. Oxford University Press**:** 427 p.

Haff L.R. (1979) An identity for the Wishart distribution with applications. *Journal of Multivariate Analysis*, **9(4):** 53-544.

Hartwig F., Dearing B.E. (1979) Exploratory data analysis. Newberry Park, CA: Sage Publications**: 711 p.**

Henderson R., Sali A., Baker M.L., Carragher B., Devkota B., Downing K.H., Egelman E.H., Feng Z., Frank J., Grigorieff N., Jiang W., Ludtke S.J., Medalia O., Penczek P.A., Rosenthal P.B., Rossmann M.G., Schmid MF., Schroeder G.F., Steven A.C., Stokes D.L., Westbrook J.D., Wriggers W., Yang H., Young J., Berman H.M., Chiu W., Kleywegt G.J., Lawson C.L. (2012) Outcome of the First Electron Microscopy Validation Task Force Meeting. *Structure,* **20 (2):** 205–214.

High R. (2000) Dealing with outliers: How to maintain your data's integrity. *University of Oregon:* http://cc.uoregon.edu/spring2000/outliers.html

Iglewicz B., Hoaglin D. (1993) How to Detect and Handle Outliers. *The ASQC Basic References in Quality Control: Statistical Techniques,* **Volume 16: 87-28.**

Leys C., Ley C., Kleina O., Bernarda P., Licataa L. (2013) Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology,* **49(4):** 764-766.

Masmaliyeva R., Murshudov G.N. (2017) Refinement and validation of macromolecular structures. *Transactions of Institute of Molecular Biology and Biotechnologies,* **1:** 80-93.

Mattson W., Rice B.M. (1999) Near-neighbor calculations using a modified cell-linked list method. *Computer Physics Communications,* **119 (2-3):** 135.

McNicholas S., Potterton E., Wilson K.S., Noble M.E.M. (2011) Presenting your structures: the CCP4mg molecular-graphics software. *Acta Cryst. D,* **67:** 386-394.

Murphy K.P. (2007) Conjugate Bayesian analysis of the Gaussian distribution. *https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf*

Nedret B., Ali S.H., Paul F.V. (2000) BACON: blocked adaptive computationally ecient outlier nominators. *Computational Statistics & Data Analysis* **34:** 279–298.

Papageorgiou A.C., Mattsson J. (2014) Protein structure validation and analysis with X-ray crystallography. *Methods Mol Biol.* **1129:**397-421.

R Core Team (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing. Vienna, Austria.* **http://www.R-project.org/**

R Core Team (2018) R: A language and environment for statistical computing. r foundation for statistical computing, Austria: Viena. ISBN 3-900051-07-0**, http://www.R-project.org**

Rupp B. (2010) Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology. Garland Science**: 800 p.**

Shiffler R.E. (1988) Maximum Z scores and outliers. *The American Statistician,* **42(No1):** 79-80.

Vagin A.A., Steiner R.S., Lebedev A.A., Potterton L., McNicholas S., Long F., Murshudov G.N. (2004) REFMAC5 dictionary: organisation of prior chemical knowledge and guidelines for its use. *Acta Cryst. D,* **60:** 2284-2295.

Venables W.N., Ripley B.D. (1999) Modern Applied Statistics with S-PLUS. Third edition. Springer: 508 p.

Winn M.D., Ballard C.C., Kevin C.D., Dodson E.J., Emsley P., Evans P.R., Keegan R.M., Krissinel E.B., Leslie A.G.W., McCoy A., McNicholas S.J., Murshudov G.N., Pannu N.S., Potterton E.A., Powell H.R., Read R.J., Vagin A., Wilsonc K.S. (2011) *Acta. Cryst.* **D67:** 235-242

Witkovsky V. (2001) Computing the distribution of a linear combination of inverted gamma variables. *Kybernetika,* **37(1):** 79-90

Atomik Temperatur Faktoru – B Qiymәti Paylanmasında Autlayerlәrin Axtarışı

R. Ç. Mәsmәliyeva[1], Q. N. Mürşüdov[1,2]

[1] AMEA Molekulyar Biologiya vә Biotexnologiyalar İnstitutunun Hesablama struktur biologiyası laboratoriyası
[2] Tibbi Tәdqiqatlar Şurasının (TTŞ) Molekulyar Biologiya İnstitutu, Kembric, Böyük Britaniya

Bu mәqalәdә atom yerdәyişmә parametrlәrindә (AYP) vә ya B qiymәtlәrinә uyğun olan ehtimal paylanmasından kәnar qiymәtlәri – autlayerlәri müәyyәn edәn bir neçә üsul bir zülal quruluşuna tәdbiq olunmuşdur. Bu mәqalәdә göstәrilir ki, belә kәnar qiymәtlәr atom modelindә olan sәhvlәri müәyyәn etmәyә kömәk edir. Bundan әlavә AYP-lәrin sürüşәn tәrs qamma paylanmasına uyğun olması hipotezi dә bir zülal tәdbiq ilә tәsdiqlәnmişdir. Biz göstәrdik ki, AYP-dә olan kәnar qiymәtlәrin aradan götürülmәsi ehtimal paylanmasının parametrlәrinin qiymәtlәrinin dәqiqliyini dә artırır. AYP-dәki lokal kәnar qiymәtlәr bu zülal quruluşunda olan sәhvlәrin harada olduğunu göstәrir. Gәlәcәkdә kәnar qiymәtlәrin tapılması nisbәtәn yaxşı zülal qurluşlarının seçilmәsinә dә kömәk edәcәk. Bundan әlavә әgәr bu proqram kristalloqrafiya vә tәk hissәcik cryo elektron mikroskopiyası vasitәsi ilә model qurulması mәrhәlәsindә istifadә edliәrsә onda alınan modelin etibarlılığı daha yüksәk olar.

Açar sözlәr: *Makromolekullar, validasiya, makro-molekulyar kristalloqrafiya, autlayet axtarışı, tәrs qamma paylanması*

Обнаружение Выброса в Атомном Температурном Факторе - Распределения Значения “B”

Р.Ч. Масмалиева [1] и Г.Н. Муршудов [1,2]

[1] Лаборатория вычислительной структурной биологии Института молекулярной биологии и биотехнологий НАН Азербайджана
[2] Институт молекулярной биологии Совета по медицинским исследованиям (*MRC*), *Кембридж, Великобритания*

Несколько методов обнаружения выбросов в параметрах атомного смещения - значения B были применены к одной конкретной макромолекулярной структуре. Показано, что выбросы в значениях “B” указывают на ошибки в атомных моделях. Этот конкретный пример доказывает верность гипотезы о скользящем обратном гамма-распределении значений “B”. Удаление выбросов из набора значений “B” улучшает оценку параметров распределения. Локальные выбросы в значениях B указывают на ошибки в атомных моделях. Нами была проверена справедливость предположения о том, что соседние атомы должны иметь одинаковые значения “B”. Ожидается, что локальная и глобальная программы обнаружения выбросов и моделирование значений “B” в качестве обратного гамма-распределения помогут выбрать надежные атомные модели. Мы предполагаем, что такое обнаружение и моделирование выбросов должно быть частью построения модели и уточнения макромолекулярных структур с использованием данных кристаллографической дифракции и карт одноэлектронной криоэлектронной микроскопии.

Ключевые слова: *Макромолекулы, валидация, макромолекулярная кристаллография, обнаружение выбросов, обратное гамма-распределение*